2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Non-stationary Bandits with Habituation and Recovery Dynamics and Knapsack Constraints

Qinyang He and Yonatan Mintz

Abstract— **Multi-armed bandit models have proven to be useful in modeling many real world problems in the areas of control and sequential decision making with partial information. However, in many scenarios, such as those prevalent in healthcare and operations management, the decision maker's expected reward will decrease if an action is selected too frequently while it may recover if they abstain from selecting this action. This scenario is further complicated when choosing a particular action also expends a random amount of a limited resource where the distribution is also initially unknown to the decision maker. In this paper we study a class of models that address this setting that we call reducing or gaining unknown efficacy bandits with stochastic knapsack constraints (ROGUEwK). We propose a combination upper confidence bound (UCB) and lower confidence bound (LCB) approximation algorithm for optimizing this model. Our algorithm chooses which action to play at each time point by solving a linear program (LP) with the UCB for the average rewards and LCB for the average costs as inputs. We show that the regret of our algorithm is sub-linear as a function of time and total constraint budget when compared to a dynamic oracle. We validate the performance of our algorithm against existing state of the art non-stationary and knapsack bandit approaches in a simulation study and show that our methods are able to on average achieve a 13% improvement in terms of total reward.**

I. INTRODUCTION

Stochastic Multi-armed bandits (MAB) have become some of the most common models for analyzing problems in sequential decision making and control with partial information. MABs have been used to model various real life applications such as medical trials [1], advertising [2], and recommendation system [3]. In the classic stochastic MAB setting, a decision maker must select an action from a finite set of actions to maximize their long term reward without knowing *a priori* the reward distribution associated with each action. This means that to be effective, their policy must balance choosing actions that may help them learn more about the system (exploration) with actions that given current information seem like they are likely to provide a high reward (exploitation). In general this reward distribution is assumed to be unchanging and stationary throughout the process and the decision maker is assumed to be able to take as many actions as they desire. However, in many real world scenarios such as those prevalent in personalized healthcare [4] and operations management [5], reward distributions may be non stationary and actions are restricted by resource constraints.

The authors are with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, 53705, USA qhe57,ymintz@wisc.edu. Authors would like to thank the support of the American Family Funding Initiative for funding this research.

Several frameworks have been used to analyze MABs with non-stationary rewards. The oldest model related to this problem is the restless bandit model proposed by Whittle [6] where bandit rewards are dependent on internal transitioning states. In more recent literature there have been two main families of non-stationary bandit models. The first family, are bandits where the non-stationary is not modeled explicitly but is subject to a total variation constraint [7]–[11]. These approaches have been shown to be effective in settings where the reward distribution either changes very slowly over time or very infrequently. The other family of models considers structured non-stationarity [12]–[14]. These approaches are more suitable for frequently changing rewards; however, they require the additional assumption that the decision maker has a model for how rewards can change over time.

Several models in the literature have been devised for settings with bandit feedback and limited resources. [15] studies the case where the cost of pulling an arm is fixed and becomes known after the arm is pulled once and [16] extends the result to the scenario where the costs are random variables. A generalization of these two settings is known as Bandit with Knapsacks (BwK) problem where each arm pull consumes multiple constrained resources. In the standard BwK setting, [17] determines the policy by solving a sequence of linear programs (LP) that take the expected reward and resource consumption as input and the optimal values of these LPs are used for the regret analysis. Recently, models have been proposed to address the BwK case where rewards can be non-stationary [11]. This model uses assumptions similar to those found in the literature that considers un-modeled non-stationarity that is bounded by a total variation constraint. This results in a dynamic regret bound that is $\mathcal{O}(\sqrt{T}\log T)$ with additional constants that depend on the total variation budget. Due to this assumption, this approach is well suited for cases where the distributions in the problem either change abruptly and infrequently or very slowly, and is not well suited for the case of frequently changing rewards. Thus, new approaches must be developed that can address both frequently changing non-stationary rewards and resource constraints.

In this paper, we propose methods to address the challenges of non-stationarity and knapsack constraints with frequently changing reward distributions. Our approach builds upon the literature related to structured non-stationarity and develops a new regret analysis for the case of dynamic regret. In particular we consider the case of non-stationarity where taking an action too frequently may reduce its reward, an effect known as habituation, while refraining from taking an action may increase its reward, known as recovery. Bandit

979-8-3503-1632-2/24/$31.00 ©2024 IEEE

models that address these effects are referred to as bandits with reducing or gaining unknown efficacy (ROGUE bandits) [12]. This form of non-stationarity is common in many real world applications such as personalized healthcare treatment [18] and online advertising [19]. In this paper we present the model we call the ROGUE bandits with knapsacks (ROGUEwK) problem and present an upper confidence bound (UCB) based approach for solving this problem we call the ROGUE knapsack UCB (ROGUEwK-UCB) algorithm. We show that our approach is able to achieve a dynamic regret bound of $\mathcal{O}(\sqrt{T}\log T)$ meeting the existing non-stationary and stationary BwK regret bound up to a log factor [11], [17]. We conduct a computational experiment and show how our approach an outperform existing approaches by 13% in terms of maximum reward.

II. Problem statement

In the ROGUEwK problem, the decision-maker is given a fixed finite set of arms $\mathcal{A}$ (with $|\mathcal{A}| = m$). At each round t, the decision maker must play one arm denoted by a_t. Each arm pull $a_t \in \mathcal{A}$ at time t provides a stochastic reward $r_{a_t,t}$ that has a sub-Gaussian distribution $\mathcal{P}_{a_t,x_{a_t,t}}$ with expectation $\mathbb{E}[r_{a_t,t}] = g(x_{a_t,t})$ for a bounded function $g_{a_t} : \mathcal{X} \to \mathbb{R}$. Each action $a \in \mathcal{A}$ has a state $x_{a,t}$ with nonlinear dynamics $x_{a,t+1} = h_a(x_{a,t}, \pi_{a,t})$ where $\pi_{i,t} = \mathbb{1}[a_t = i], h : \mathcal{X}\times\mathbb{B} \to \mathcal{X}$ is a known dynamics function, and $\mathcal{X}$ is a compact convex set such that $x_{a,t} \in \mathcal{X}\ \forall a, t$ and $x_{a,0}$ is initially unknown for $a \in \mathcal{A}$. The maximum time horizon T is finite and known in advance. We also assume that there are d resources that each has a budget B_j. Without loss of generality, we assume $B_j = B$ for all j. Each arm pull $a_t \in \mathcal{A}$ at time t incurs a stochastic consumption $c_{a_t,j,t}$ on resource j with support in $[0, 1]$ and denote the expected consumption matrix to be $\mathbf{C}$ where $\mathbf{C}_{ij}$ denotes the expected consumption on resource j for arm i. The realizations of consumption $c_{a,j,t}$ are independently and identically distributed and in each round the resource consumption of the pulled arm is revealed to the decision maker. The interaction between the learner and the environment terminates at the earliest time τ when at least one constraint is violated, i.e. $\sum_{t=1}^{\tau} c_{a_t,j,t} > B$, or the time horizon T is exceeded. The objective is to maximize the expected cumulative reward until time τ, i.e. $\mathbb{E}\left[\sum_{t=1}^{\tau-1} r_{a_t,t}\right]$. We measure the performance of algorithm/policy Π by its regret which is defined as: $\mathrm{Reg}(\Pi, T) := \mathrm{OPT}(T) - \mathbb{E}\left[\sum_{t=1}^{\tau-1} r_{a_t,t} \mid \Pi\right]$. Here $\mathrm{OPT}(T)$ denotes the expected cumulative reward of the optimal dynamic policy given all the information on the initial state, reward and cost distributions.

A. Technical Assumptions on ROGUEwK

As part of our analysis we make the following technical assumptions. First is a set of assumptions introduced in [12] for the analysis of bandits with ROGUE non-stationarity.

Assumption 1. *$r_{a,t}$ are conditionally independent given $x_{a,0}$ (or equivalently, the complete sequence of $x_{a,t}, a_t$).*

This is a fairly mild assumption that is a non-stationary analogue to the classical MAB assumption of i.i.d rewards. Essentially it implies that at any two points in time t, t' such that $t \neq t'$, $r_{a,t} \mid \{x_{a,t},\}$ is independent of $r_{a,t'} \mid \{x_{a,t'},\}$.

Assumption 2. *For all $a \in \mathcal{A}$, the reward distribution $\mathcal{P}_{a,x}$ has a log-concave probability density function (p.d.f.) $p_a(r \mid x)$ for all $x \in \mathcal{X}$.*

This assumption provides regularity for the reward distributions and many common distributions (e.g., Gaussian and Bernoulli) have this property. Let $f(\cdot)$ be L-Lipschitz continuous if $|f(x_1) - f(x_2)| \leq L \|x_1 - x_2\|_2$ for all x_1, x_2 in the domain of f, and let f to be locally L-Lipschitz on a compact set $\mathcal{S}$ if it has the Lipschitz property for all points of its domain on set $\mathcal{S}$. Our next assumption is on the stability of the above distributions with respect to various parameters.

Assumption 3. *The log-likelihood ratio $\ell(r; x', x) = \log(p(r \mid x')/p(r \mid x))$ of the distribution family $\mathcal{P}_{a,x}$ is locally L_f-Lipschitz with respect to x on the compact set $\mathcal{X}$ for all values of $x' \in \mathcal{X}$, and g is locally L_g-Lipschitz with respect to x on the compact set $\mathcal{X}$.*

This assumption guarantees that when two sets of parameters have similar values, the resulting distributions will be close to each other. We also introduce an additional assumption regarding the functional form of the reward distribution family.

Assumption 4. *The reward distribution $\mathcal{P}_{a,x_a}$ for all $x_a \in \mathcal{X}$ and $a \in \mathcal{A}$ is sub-Gaussian with parameter σ, and either $p(r \mid x)$ has a finite support or $\ell(r; x', x)$ is locally L_p-Lipschitz with respect to r.*

This assumption is essential to guarantee that sample averages closely approximate their means, and it is upheld by various distributions (such as a Gaussian location family with a known variance). We impose the following conditions regarding the dynamics governing the state of each action.

Assumption 5. *The dynamic transition function h is bijective and L_h-Lipschitz continuous such that $L_h < 1$.*

This assumption ensures that there are no rapid changes in the states of each action, and implies stability in the dynamics. The last assumption, drawn from existing BwK literature, pertains to the total budget constraint of the problem.

Assumption 6 (Linear Growth)**.** *The resource budget $B = bT$ for some $b > 0$.*

This assumption is needed to quantify the relationship between B and T and this is a common assumption used in BwK literature [11], [20].

B. Preliminaries: Concentration Inequalities

Before diving into the details of the ROGUEwK-UCB, we introduce some preliminary results that we use in our analysis. UCB algorithms have been commonly used for stochastic non-stationary MAB [11], [12], [21]. To construct the UCB and LCB we need to use appropriate concentration inequalities for the parameter estimates of the rewards and costs. To analyze

the concentration of the cost parameter estimates we use the Azuma-Hoeffding Inequality in the following form:

Lemma 1 (Azuma-Hoeffding's Inequality [22])**.** *Consider a random variable with distribution supported on* $[0,1]$. *Denote its expectation as* z. *Let* $\bar{Z}$ *be the average of* N *independent samples from this distribution. Then,* $\forall \delta > 0$, *the following inequality holds with probability at least* $1-\delta$, $|\bar{Z} - z| \leq \sqrt{\frac{1}{2N}\log(\frac{2}{\delta})}$. *More generally, this result holds if* $Z_1, \ldots, Z_N \in [0,1]$ *are random variables,* $\bar{Z} = \frac{1}{N}\sum_{n=1}^{N} Z_n$, *and* $z = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[Z_n \mid Z_1, \ldots, Z_{n-1}\right]$.

For analyzing rewards, a limitation of this inequality is that it requires the random variables to be independent and in the context of learning, requires the use of unbiased estimators. These conditions are violated in the case of ROGUE rewards as due to the structure of the model maximum likelihood estimators (MLE), which are in general biased, will be more effective then unbiased ones. This necessitates the use of an alternative concentration inequality. The key of this concentration has to do with a quantity called the trajectory Kullbek-Liebler (KL) divergence that is defined as follows:

Definition 1 (Definition 1 from [12])**.** For some input action sequence π_1^T and arm $a \in \mathcal{A}$ with dynamics h_a, given starting parameter values $x_{a,0} \in \mathcal{X}$, let $\mathcal{T}_a(T) \subset \{1, ..., T\}$ be the set of times when action a was chosen up to time T, then define the trajectory KL-divergence between these two trajectories with the same input sequence and different starting conditions as: $D_{a,\pi_1^T}(x_{a,0} \| x'_{a,0}) = \sum_{t \in \mathcal{T}_a(T)} D_{KL}(\mathcal{P}_{a,x_{a,t}} \| \mathcal{P}_{a,x'_{a,t}}) = \sum_{t \in \mathcal{T}_a(T)} D_{KL}(\mathcal{P}_{a,h_a^t(x_{a,0})} \| \mathcal{P}_{a,h_a^t(x'_{a,0})})$.

Where h_a^k represents the functional composition of h_a with itself k times subject to the given input sequence, $\mathcal{P}_{a,x}$ is the probability law of the system under parameters x, and D_{KL} is the standard KL divergence. Using this quantity we use the following concentration result:

Theorem 1 (Theorem 1 from [12])**.** *Let* $x^*_{a,0}$ *be the true initial state of an arm and let* $\hat{x}_{a,0}$ *be the MLE estimate for this parameter. That is,* $\hat{x}_{a,0} = \arg\min\{-\sum_{t \in \mathcal{T}_a(T)} \log p(r_t \mid x_{a,t}) : x_{a,t+1}. = h_a(x_{a,t}, \pi_{a,t})\}$, *where* $\{r_t\}_{t \in \mathcal{T}_a}$ *are the observed rewards for action* $a \in \mathcal{A}$. *Let* $n_a(T) = |\mathcal{T}_a|$ *denote the number of times arm* a *is played up to time* T. *Then for* $\alpha \in (0,1)$, *with probability at least* $1-\alpha$, *we have* $\frac{1}{n_a(T)} D_{a,\pi_1^T}(x^*_{a,0} \| \hat{x}_{a,0}) \leq B(\alpha)\sqrt{\frac{\log(1/\alpha)}{n_a(T)}}$, *where* $B(\alpha) = \frac{c_f(d_x)}{\sqrt{\log(1/\alpha)}} + L_p\sigma\sqrt{2}$ *and* $c_f(d_x) = 8L_f \operatorname{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{a_x}} L_f \cdot \operatorname{diam}(\mathcal{X})\sqrt{\pi d_x}$.

Here the inclusions of the terms related to $B(\alpha)$ account for the bias in the MLE estimation.

III. ROGUEwK-UCB Algorithm

In this section, we explain the details of the ROGUEwK-UCB algorithm. For the stationary BwK, the optimal dynamic policy can be computed by solving a LP that takes the mean reward and mean consumption vectors as input [11], [20]. Also, the expected cumulative reward of this optimal dynamic policy is used for regret analysis. Similarly we introduce a nonlinear optimization that upper bounds the expected reward of the optimal dynamic policy in non-stationary case and use it to construct the ROGUwK-UCB algorithm.

A. Relation of single step and multi-step problems

Let $\mathbf{x}_t = (x_{1,t}, x_{2,t}, ..., x_{m,t})$, $\mathbf{g}_t = (g_1(x_{1,t}), g_2(x_{2,t}), ..., g_1(x_{m,1}))$. Define $NLP(\mathbf{x}_0, T, \mathbf{C})$ as $\max_{\boldsymbol{\pi}_t} \left\{ \sum_{t=1}^{T} \boldsymbol{\pi}_t^\top \mathbf{g}_t : x_{a,t} = h_a(\boldsymbol{\pi}_{a,t}, x_{a,t}) \ \forall a, t \in \mathcal{A} \times \{1, ..., T\}, \ \sum_{t=1}^{T} \mathbf{c}_j^\top \boldsymbol{\pi}_t \leq B \ \forall j \in \{1, ..., d\}, \ \boldsymbol{\pi}_t \in \Delta_m \ \forall t \in \{1, ..., T\} \right\}$. Here $\Delta_m \in \mathbb{R}^m$ is the m-dimensional unit simplex and $\mathbf{c}_j$ is the jth column of the matrix $\mathbf{C}$. Notice that the optimal value of $NLP(\mathbf{x}_0, T, \mathbf{C})$ is an upper bound on the expected cumulative reward of the optimal dynamic policy because it is the linear relaxation of the actual decision making problem which requires all the variables to be binary. This nonlinear optimization is hard to solve without specific structure in costs and state transition dynamics. A similar problem where the rewards have no state dependency has proven to be PSPACE-hard [23]. We instead consider solving a LP to make step-wise decisions. Define the single-step optimization problem at time t $LP(\mathbf{g}_t, \mathbf{C})$ with respect to $\{x_{a,t}\}_{a \in \mathcal{A}}$, to be: $\max_{\boldsymbol{\pi}_t} \left\{ \boldsymbol{\pi}_t^\top \mathbf{g}_t : \sum_{\mathcal{A}} \mathbf{c}_j^\top \boldsymbol{\pi}_t \leq b \ \forall j \in \{1, ..., d\}, \ \boldsymbol{\pi}_t \in \Delta_m \right\}$. The single-step LP problem can be interpreted as determining the optimal pulling distribution under a normalized resource budget b. The following proposition establishes the relationship between the global nonlinear optimization problem and step-wise linear program.

Proposition 1. $NLP(\mathbf{x}_0, T, \mathbf{C}) \leq T \cdot LP(\mathbf{g}_0, \mathbf{C}) + L_g \frac{1}{1-L_h} \operatorname{diam}(\mathcal{X})$

Proof. Denote the optimal solution for $NLP(\mathbf{x_0}, T, \mathbf{C})$ as $\{\boldsymbol{\pi}_t^{NLP}\}_{t=1}^T$ and the optimal solution for $LP(\mathbf{g}_0, \mathbf{C})$ as π_0^*. $NLP(\mathbf{x}_0, T, \mathbf{C}) - T \cdot LP(\mathbf{x}_0, \mathbf{C}) = \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_{a,t}^{NLP} g(h^t(x_{a,0}) - \pi^*_{a,0} g(x_{a,0}) \leq \sum_{t=1}^{T} \max_{a \in \mathcal{A}} g(h^t(x_{a,0})) - g(x_{a,0}) \leq \sum_{t=1}^{T} L_g \max_{a \in \mathcal{A}} ||h^t(x_{a,0}) - x_{a,0}||_2 \leq \sum_{t=1}^{T} L_g \max_{a \in \mathcal{A}} ||h^t(x_{a,0}) - h^t(x')||_2 \leq \sum_{t=1}^{T} L_g L_h^t \operatorname{diam}(\mathcal{X}) \leq L_g \frac{1}{1-L_h} \operatorname{diam}(\mathcal{X})$. Inequality 5d-5e comes from the fact that h is a bijective function. $x_{a,0} = h(x') = h^2(x'') = ... = h^t(x''')$ for some $x', x'', x''' \in \mathcal{X}$. ∎

B. Algorithm Details

Next, we present the ROGUEwK-UCB algorithm as shown in Algorithm 1. We initialize the algorithm by pulling each arm once. After the first m rounds, in every time step t, we first compute the MLE estimates of each arm's initial states and calculate the UCB for rewards based on the estimates of the states using Theorem 1. The idea behind the UCB is as follows: from Theorem 1, we know that with high probability, the true initial states are within a certain trajectory divergence from the MLE estimates, so we find the largest possible value of $g(x_{a,t})$ within the designated confidence radius. We also

compute lower confidence bounds (LCB) for costs of each arm pull on different resources. Then we solve a single step LP problem which takes the UCBs and LCBs as input. The optimal solution to this LP is the probability distribution according to which the arm is going to be played and we pick an arm randomly following this distribution.

Remark 1. While Algorithm 1 is written such that it requires knowledge of time horizon T, if it is to be run indefinitely one could use the doubling trick [24] and still preserve its statistical properties.

Algorithm 1 ROGUEwK-UCB

Require: Transition function $\{h_a\}$, reward function $\{g_a\}$

for $t \leq |\mathcal{A}|$ **do**
  Pick an arm a that hasn't been chosen before
end for
for $|\mathcal{A}| \leq t \leq T$ **do**
  for $a \in \mathcal{A}$ **do**
    Compute
    $(\hat{x}_{a,0}) = \arg\min \left\{ - \sum_{s \in \mathcal{T}_a(t)} \log p(r_s \mid x_{a,s}) : x_{a,s+1} = h_a(x_{a,s}, \pi_{a,s}) \text{ for } t \in \{0, \dots, T\} \right\}$
    $g_{a,t}^{UCB} = \max_{x_{a,0} \in \times \mathcal{X}} \{ g(h_a^t(x_{a,0})) : \frac{1}{n_a(t)} D_{a,\pi_a^t}(x_{a,0} \| \hat{x}_{a,0}) \leq B(6mT^2) \sqrt{\frac{\log(6mT^2)}{n_{a_s}(t)}} \}$
    $c_{a,j,t}^{LCB} = \hat{c}_{a,j,n_a(t)} - \sqrt{\frac{1}{2n_a(t)} \log(12mdT^2)}$ $\forall j \in \{1, ..., d\}$ where $\hat{c}_{a,j,n_a(t)} = \frac{1}{n_a(t)} \sum_{s \in \mathcal{T}_a(t)} c_{j,s}$
  end for
  Solve the single-step problem $LP(g_t^{UCB}, \mathbf{C}^{LCB})$ and denote its optimal solution by $\boldsymbol{\pi}_t^* = (\pi_{1,t}^*, ..., \pi_{m,t}^*)$
  Pick arm a_t randomly according to $\boldsymbol{\pi}_t^*$, i.e., $\mathbb{P}(a_t = i) = \pi_{i,t}^*$
  Observe reward r_t and consumption $c_{j,t}, \forall j$.
  Terminate if budget is exceeded
end for

IV. Regret Analysis for ROGUEwK-UCB

In the following section, we present the analysis for the regret of the ROGUEwK-UCB algorithm. We first bound the difference between the maximum time horizon and the termination time when budget is exhausted.

Proposition 2. *The following inequality holds with probability at least* $1 - \frac{1}{2T}$*:* $T - \tau \leq \frac{1}{b}(1 + 4\sqrt{m} + \sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)})$.

Proof. From Hoeffding's inequality [25] we have that for any $a \in \mathcal{A}$, $j \in 1, ..., d$, $t \leq \min\{\tau, T\}$, with probability at least $1 - \frac{1}{6mdT^2}$, $|\hat{c}_{a,j,t} - \mathbf{C}_{a,j}| \leq \sqrt{\frac{1}{2n_{a_t}(t)} \log(12mdT^2)}$. Then: $|\mathbf{C}_{a_t,j} - c_{a_t,j,t}^{LCB}| = |\mathbf{C}_{a_t,j} - (\hat{c}_{a_t,j,t} - \sqrt{\frac{1}{2n_{a_t}(t)} \log(12mdT^2)})| \leq |\mathbf{C}_{a_t,j} - \hat{c}_{a_t,j,t}| + \sqrt{\frac{1}{2n_{a_t}(t)} \log(12mdT^2)} \leq 2\sqrt{\frac{1}{2n_{a_t}(t)} \log(12mdT^2)}$.

Then for all $t \leq \min\{\tau, T\}$ with probability at least $1 - T \cdot m \cdot d \cdot \frac{1}{6mdT^2} = 1 - \frac{1}{6T}$: $|\sum_{s=1}^t (\mathbf{C}_{a_s,j} - c_{a_s,j,s}^{LCB})| \leq \sum_{s=1}^t |\mathbf{C}_{a_s,j} - c_{a_s,j,s}^{LCB}| \leq \sum_{s=1}^t 2\sqrt{\frac{1}{2n_{a_s}(s)} \log(12mdT^2)} = \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{T}_a(t)} 2\sqrt{\frac{1}{2n_{a_s}(s)} \log(12mdT^2)} = \sum_{a \in \mathcal{A}} \sum_{s=1}^{|\mathcal{T}_a(t)|} 2\sqrt{\frac{1}{2s} \log(12mdT^2)} \overset{(a)}{\leq} \sum_{a \in \mathcal{A}} 4\sqrt{\frac{|\mathcal{T}_a(t)|}{2} \log(12mdT^2)} \overset{(b)}{\leq} 4\sqrt{\frac{mT}{2} \log(12mdT^2)}$ where (a) comes from the fact that $\sum_{n=1}^N \frac{1}{\sqrt{n}} \leq 2\sqrt{N}$ and (b) follows from the Cauchy-Shwartz inequality. By Lemma 1 with probability at least $1 - \frac{1}{6T} \leq 1 - \frac{1}{6mdT^2}$: $|\sum_{s=1}^t c_{j,s} - \mathbf{C}_{a_s,j}| \leq \sqrt{\frac{t}{2}\log(12mdT^2)} \leq \sqrt{\frac{T}{2}\log(12mdT^2)}$. Denote $\mathbf{c}_{j,t}^{LCB} = (c_{1,j,t}^{LCB}, ..., c_{m,j,t}^{LCB})$. Note that $\mathbb{E}[\mathbf{c}_{j,t}^{LCB^\top} \boldsymbol{\pi}_t^*] = \mathbb{E}[c_{j,a_t,t}^{LCB}]$ and $c_{a_t,j,t}^{LCB} \in [-\sqrt{\frac{1}{2}\log(12mdT^2)}, 1]$, then by Lemma 1 we have with probability at least $1 - \frac{1}{6T} \leq 1 - \frac{1}{6mdT^2}$: $|\sum_{s=1}^t \mathbf{c}_{j,s}^{LCB^\top} \boldsymbol{\pi}_s^* - c_{a_s,j,s}^{LCB}| \leq (1 + \sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)})$. Combining the above with probability at least $1 - 3 \cdot \frac{1}{6T} = 1 - \frac{1}{2T}$: $|\sum_{s=1}^t c_{j,s} - \mathbf{c}_{j,s}^{LCB^\top} \boldsymbol{\pi}_s^*| \leq |\sum_{s=1}^t c_{j,s} - \mathbf{C}_{a_s,j}| + |\sum_{s=1}^t (\mathbf{C}_{a_s,j} - c_{a_s,j,s}^{LCB})| + |\sum_{s=1}^t c_{a_s,j,s}^{LCB} - \mathbf{c}_{j,s}^{LCB^\top} \boldsymbol{\pi}_s^*| \leq (1 + 4\sqrt{m} + \sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)})$. Without loss of generality, we analyze the case when $\tau \leq T$. At termination time τ, let $c_{j,t}$ denote the realized cost, then $\sum_{t=1}^\tau c_{j,t} \geq bT$ for some $j \leq d$. From the fact that for all time t, $\boldsymbol{\pi}_t^*$ is a feasible solution to the problem $LP(g_t^{UCB}, \mathbf{C}^{LCB})$, we have $\sum_{t=1}^\tau \mathbf{c}_{j,t}^{LCB^\top} \boldsymbol{\pi}_t^* \leq b\tau$. Combining this inequality with the previous inequality, we have with probability at least $1 - \frac{1}{2T}$: $\sum_{t=1}^\tau c_{j,t} \leq b\tau + (1 + 4\sqrt{m} + \sqrt{\frac{1}{2}\log(12mdT^2)}) \cdot (\sqrt{\frac{T}{2}\log(12mdT^2)})$. Therefore we have $(1+4\sqrt{m}+\sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)}) \geq b(T - \tau)$, which yields the desired result. ∎

Next we bound the absolute difference of cumulative realized rewards and optimal values of single step LPs.

Proposition 3. *The following inequality holds for all* $t \leq \min\{\tau, T\}$*with probability at least* $1 - \frac{1}{2T}$*:* $|\sum_{s=1}^t r_s - g_s^{UCB^\top} \boldsymbol{\pi}_s^*| \leq \sqrt{2T\sigma^2 \log(12T)} + (\frac{1}{1-L_h} + \sqrt{\frac{T}{2}\log(12T)}) L_g \operatorname{diam}(\mathcal{X})$.

Proof. From the Hoeffding's inequality for sum of independent Sub-Gaussian random variables [25], we have that with probability at least $1 - \frac{1}{6T}$: $|\frac{1}{t}\sum_{s=1}^t (r_s - g_{a_s})| \leq \sqrt{\frac{2}{t}\sigma^2 \log(12T)} \implies |\sum_{s=1}^t (r_s - g_{a_s})| \leq \sqrt{2t\sigma^2 \log(12T)} \leq \sqrt{2T\sigma^2 \log(12T)}$. Note that for any s, $\mathbf{E}[\boldsymbol{\pi}_s^{*^\top} \boldsymbol{g}_s^{UCB}] = \mathbf{E}[g_{a_s}^{UCB}]$, by Lemma 1, we have with probability $1 - \frac{1}{6T}$: $|\sum_{s=1}^t x_s^{*^\top} g_s^{UCB} - g_{a_s}^{UCB}| \leq L_g \operatorname{diam}(\mathcal{X}) \sqrt{\frac{T}{2}\log(12T)}$. We denote $x_{a_s,s}^{UCB} = \arg\max_{x_{a,0} \in \mathcal{X}} \{ g(h_a^t(x_{a,0})) : \frac{1}{n_a(t)} D_{a,\pi_1^t}(x_{a,0} \| \hat{x}_{a,0}) \leq B(6mT^2) \sqrt{\frac{\log(6mT^2)}{n_{a_s}(t)}} \}$ By Theorem 1 we have with probability at least $1 - T \times m \times \frac{1}{6mT^2} = 1 - \frac{1}{6T}$ the following in-

equality holds: $|\sum_{s=1}^{t}(g_{a_s}-g_{a_s}^{UCB})| \le \sum_{s=1}^{t}|g(h^s(x_{a_s}))-g_{a_s}(h^s(x_{a_s,s}^{UCB}))| \le \sum_{s=1}^{t} L_g|h^s(x_{a_s,s})-h^s(x_{a_s,s}^{UCB})| \le \sum_{t=1}^{T} L_g L_h^t \operatorname{diam}(\mathcal{X}) \le L_g \frac{1}{1-L_h}\operatorname{diam}(\mathcal{X})$ Then combining the above with the previous results with probability at least $1-3\times\frac{1}{6T}=1-\frac{1}{2T}$, $|\sum_{s=1}^{t} r_s - g_s^{UCB^\top}\boldsymbol{\pi}_s^*| \le |\sum_{s=1}^{t}(r_s-g_{a_s})|+|\sum_{s=1}^{t}(g_{a_s}-g_{a_s}^{UCB})|+|g_{a_s}^{UCB}-\sum_{s=1}^{t}\boldsymbol{\pi}_s^{*\top} g_s^{UCB}| \le \left(\frac{1}{1-L_h}+\sqrt{\frac{T}{2}\log(12T)}\right)L_g\operatorname{diam}(\mathcal{X})+\sqrt{2T\sigma^2\log(12T)}$. ■ Combining Propositions 1,2,3 we can derive the final result.

Theorem 2. *Under Assumption 1–6, the regret of Algorithm 1 is upper bounded as* $\operatorname{Reg}(\Pi^{\textit{ROGUEwK-UCB}}, T) \le \frac{1}{b}(1+4\sqrt{m}+\sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)})\cdot LP(\mathbf{g}_0,\mathbf{C}) + (\frac{2}{1-L_h}+\sqrt{\frac{T}{2}\log(12T)})L_g\operatorname{diam}(\mathcal{X}) + \sqrt{2T\sigma^2\log(12T)} + LP(\mathbf{g}_0,\mathbf{C}) = \mathcal{O}(\frac{1}{b}\sqrt{mT}\log(mdT))$.

Proof. Let $\boldsymbol{\pi}_0^{LP}$ denote the optimal solution to $LP(\mathbf{g}_0,\mathbf{C})$: $\sum_{t=1}^{\tau-1} \boldsymbol{g}_t^{UCB^\top}\boldsymbol{\pi}_t^* - (\tau-1)LP(\mathbf{g}_0,\mathbf{C}) = \sum_{t=1}^{\tau-1}(\boldsymbol{g}_t^{UCB^\top}\boldsymbol{\pi}_t^* - \boldsymbol{g}_0^\top\boldsymbol{\pi}_0^{LP}) \le \sum_{t=1}^{\tau-1}\max_{a\in\mathcal{A}} g(h^t(x_{a,0})) - g(x_{a,0}) \le \sum_{t=1}^{\tau-1} L_g \max_{a\in\mathcal{A}}||h^t(x_{a,0}) - x_{a,0}||_2 \le \sum_{t=1}^{\tau-1} L_g \max_{a\in\mathcal{A}}||h^t(x_{a,0}) - h^t(x')||_2 \le \sum_{t=1}^{\tau-1} L_g L_h^t \operatorname{diam}(\mathcal{X}) \le \frac{L_g}{1-L_h}\operatorname{diam}(\mathcal{X})$. Then by Propositions 1,2,3 we have probability at least $1-\frac{1}{T}$: $OPT - \sum_{t=1}^{\tau-1} r_t \le NLP(\mathbf{x}_0,T,\mathbf{C}) - \sum_{t=1}^{\tau-1} r_t = (NLP(\mathbf{x}_0,T,\mathbf{C}) - \sum_{t=1}^{\tau-1}\boldsymbol{g}_t^{UCB^\top}\boldsymbol{\pi}_t^*) + (\sum_{t=1}^{\tau-1}\boldsymbol{g}_t^{UCB^\top}\boldsymbol{\pi}_t^* - \sum_{t=1}^{\tau-1} r_t) \le (T-\tau+1)LP(\mathbf{g}_0,\mathbf{C}) + L_g\frac{1}{1-L_h}\operatorname{diam}(\mathcal{X}) + (\frac{1}{1-L_h}+\sqrt{\frac{T}{2}\log(12T)})L_g\operatorname{diam}(\mathcal{X}) + \sqrt{2T\sigma^2\log(12T)} \le \frac{1}{b}(1+4\sqrt{m}+\sqrt{\frac{1}{2}\log(12mdT^2)})(\sqrt{\frac{T}{2}\log(12mdT^2)})\cdot LP(\mathbf{g}_0,\mathbf{C}) + (\frac{2}{1-L_h}+\sqrt{\frac{T}{2}\log(12T)})L_g\operatorname{diam}(\mathcal{X}) + \sqrt{2T\sigma^2\log(12T)} + LP(\mathbf{g}_0,\mathbf{C}) = \mathcal{O}(\frac{1}{b}\sqrt{mT}\log(mdT))$. Note that $\operatorname{OPT}(T)$ is of linear T ($\operatorname{OPT}(T) \le T\cdot L_g\operatorname{diam}(\mathcal{X})$) that transforms the high probability bound into the expectation bound. ■

Remark 2. This result is significant because $\mathcal{O}(\frac{1}{b}\sqrt{mT}\log(mdT))$ is sublinear in T, and for fixed T, it is also sublinear in the total budget B based on the relationship between T and B introduced in Assumption 6. Regarding other state of the art results, our results matches the $\Omega(\sqrt{mOPT}\log(T))$ for stochastic BwK setting in [17] given that $OPT=\Theta(T)$ up to several log factors. However it is important to note that unlike the stationary setting, our result contains constants that are dependant on the non-stationarity of the system. Our result also matches [11] where the non-stationarity of BwK is defined by the global non-stationarity budget.

V. Numerical Experiments

In this section, we perform numerical experiments to demonstrate the effectiveness of ROGUEwK-UCB. We consider a dynamic generalized linear model (GLM) [26], [27] which can be interpreted as non-stationary generalizations of the classical (Bernoulli reward) stationary MAB [28]–[30]. The exact dynamics are as follows: the transition function $h(x_t,\pi_t) = A_a x_t + B_a\pi_t + K_a$ where A_a, B_a, K_a are matrices/vectors of the correct size and and the rewards $r_{a,t}$ are Bernoulli with a logistic link function of the form $\mathbb{E}[r_{a,t}] = g_a(x_{a,t}) = \frac{1}{1+\exp(-\alpha_a-\beta_a^\top x_{a,t})}$ where α is the vector of the correct size and $\beta_a \in \mathcal{R}$. We assume there are three resource constraints and the consumption distribution is uniform. We include three arms in this experiment whose dynamics and support of consumption distribution are shown in Table I and Table II. Arm 1 has small habituation and recovery effects and is stable in its reward (indicated by high k and β) but it has an unproportionally high cost for one of the resources. Arm 2 has moderate habituation and recovery effects and moderate consumption while Arm 3 has strong habituation and recovery effects and on average less consumption than Arm 2.

We compare our ROGUEwK-UCB algorithm against two other algorithms: the first is the naive UCB algorithm [21] that does not take non-stationarity and resource consumption into consideration; the second is the sliding window upper confidence bounds algorithm (SW-UCB) [11] that can handle non-stationary in both rewards and costs by using a sliding window on the UCB estimates. For both algorithms we used the theoretically optiamlly derived hyper-parameters [11]. We set the maximum time horizon T to be 1,000 and test the cumulative rewards for budget from 10 to 300. Each of the candidate algorithms was replicated 10 times.

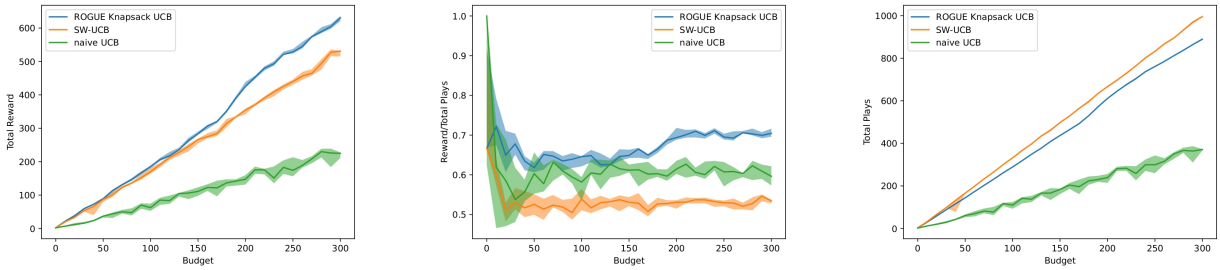
Figures 1a,1b, and 1b show the cumulative reward collected by all algorithms within the maximum allowed time horizon, the average reward per play for each algorithm, and the total number of plays for each algorithms respectively. The solid line represent the median value across 10 replicates and the shaded area represents the interquartile range of the values among 10 replicates. As depicted in the plots, ROGUEwK-UCB achieves the most total reward across all budgets. Compared with naive UCB, both BwK algorithms are cost-aware and avoid exhausting budget early by picking costly arms and thus achieve higher cumulative reward. Compared with SW-UCB, although ROGUEwK-UCB has fewer plays before the budget is exhausted, it picks more cost-effective arms since it estimates reward based information of underlying non-stationary dynamics. ROGUEwK-UCB gains on average 13% more total reward than SW-UCB across all budgets.

TABLE I: Parameters for the dynamics of each Arm

Action	x_0	A	B	K	α	β
0	0.1	0.2	-0.5	0.8	0.2	0.8
1	0.3	0.7	-1.2	0.4	0.5	0.3
2	0.9	0.5	-2.0	1.0	0.1	1.0

TABLE II: Support for the cost of each arm and resource

Arm \ Resource	1	2	3
0	[0.1,0.2]	[0.6,0.8]	[0.3,0.5]
1	[0.2,0.3]	[0.3,0.4]	[0.1,0.5]
2	[0.2,0.3]	[0.2,0.4]	[0.1,0.3]



(a) (b) (c)

Fig. 1: Cumulative reward 1a, average reward 1b, and total plays 1c for each algorithm by budget.

VI. Conclusion

We investigated non-stationary bandits with reducing or gaining unknown efficacy and knapsack constraints. We proposed an efficient UCB algorithm that determines the arms to play by solving a LP taking the UCB estimates of the rewards and LCB estimates of the costs as input. We showed that this algorithm achieves sublinear regret in terms of time horizon compared to a dynamic oracle. Numerical experiments demonstrated that our algorithm outperforms other state-of-art algorithms for non-stationary BwK.

References

[1] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Statistical science: a review journal of the Institute of Mathematical Statistics*, vol. 30, no. 2, p. 199, 2015.

[2] E. M. Schwartz, E. T. Bradlow, and P. S. Fader, "Customer acquisition via display advertising using multi-armed bandit experiments," *Marketing Science*, vol. 36, no. 4, pp. 500–522, 2017.

[3] C. Zeng, Q. Wang, S. Mokhtari, and T. Li, "Online context-aware recommendation with time varying multi-armed bandit," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 2025–2034.

[4] T. Zhou, Y. Wang, L. Yan, and Y. Tan, "Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach," *Information Systems Research*, vol. 34, no. 4, pp. 1493–1512, 2023.

[5] S. Krishnasamy, R. Sen, R. Johari, and S. Shakkottai, "Learning unknown service rates in queues: A multiarmed bandit approach," *Operations research*, vol. 69, no. 1, pp. 315–330, 2021.

[6] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

[7] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *arXiv preprint arXiv:0805.3415*, 2008.

[8] J. Y. Yu and S. Mannor, "Piecewise-stationary bandit problems with side observations," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1177–1184.

[9] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," *Advances in neural information processing systems*, vol. 27, 2014.

[10] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Hedging the drift: Learning to optimize under nonstationarity," *Management Science*, vol. 68, no. 3, pp. 1696–1713, 2022.

[11] S. Liu, J. Jiang, and X. Li, "Non-stationary bandits with knapsacks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 522–16 532, 2022.

[12] Y. Mintz, A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka, "Non-stationary bandits with habituation and recovery dynamics," *Operations Research*, vol. 68, no. 5, pp. 1493–1516, 2020.

[13] J. Gornet, M. Hosseinzadeh, and B. Sinopoli, "Stochastic multi-armed bandits with non-stationary rewards generated by a linear dynamical system," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 1460–1465.

[14] L. Wei and V. Srivastava, "On distributed multi-player multiarmed bandit problems in abruptly changing environment," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 5783–5788.

[15] L. Tran-Thanh, A. Chapman, A. Rogers, and N. Jennings, "Knapsack based optimal policies for budget–limited multi–armed bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 1134–1140.

[16] W. Ding, T. Qin, X.-D. Zhang, and T.-Y. Liu, "Multi-armed bandit with budget constraint and variable costs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 232–238.

[17] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, "Bandits with knapsacks," *Journal of the ACM (JACM)*, vol. 65, no. 3, pp. 1–55, 2018.

[18] P. Liao, K. Greenewald, P. Klasnja, and S. Murphy, "Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–22, 2020.

[19] D. Bertsimas and A. J. Mersereau, "A learning approach for interactive marketing to a customer segment," *Operations Research*, vol. 55, no. 6, pp. 1120–1135, 2007.

[20] N. Immorlica, K. Sankararaman, R. Schapire, and A. Slivkins, "Adversarial bandits with knapsacks," *Journal of the ACM*, vol. 69, no. 6, pp. 1–47, 2022.

[21] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, pp. 235–256, 2002.

[22] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal, Second Series*, vol. 19, no. 3, pp. 357–367, 1967.

[23] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," in *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 1994, pp. 318–322.

[24] L. Besson and E. Kaufmann, "What doubling tricks can and can't do for multi-armed bandits," *arXiv preprint arXiv:1803.06971*, 2018.

[25] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

[26] P. McCullagh, *Generalized linear models*. Routledge, 2019.

[27] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, "Parametric bandits: The generalized linear case," *Advances in neural information processing systems*, vol. 23, 2010.

[28] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 41, no. 2, pp. 148–164, 1979.

[29] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[30] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual conference on learning theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 359–376.